# Projet ANR Acceler-AI

## Adaptive Co-Construction of Ethics for LifElong tRustworthy AI

---

Rémy Chaput

2023/03/13

Rencontre du GT ACE

# Contexte

- Impact des systèmes d'IA dans la société
- Ces systèmes doivent être alignés avec les valeurs morales humaines
- Prise en compte de l'éthique à divers stades de la conception (*In / For / By Design*)
- Critères attendus :
  - **Diversité** de valeurs, d'acteurs, de situations
  - **Longue durée** des systèmes, et évolution continue de leur environnement
  - **Fiabilité** des systèmes pour être acceptés par les utilisateurs

# Objectif général

Co-construction adaptative de l’éthique, dans un système intelligent socio-technique, gérant la diversité, le long terme, et la fiabilité

# Challenges

- **IA centrée sur l'humain** : comment le système peut accomplir ses tâches en étant aligné avec les valeurs morales humaines
- **IA sûre** : comment s'assurer que le système ne sorte pas des bornes spécifiées
- **IA adaptative** : comment le système peut s'adapter sur le plan technique et sociétal

# Hypothèses

1. Modèle éthique résulte d'un processus de co-construction centré sur l'humain, et d'un apprentissage en situation
2. Interaction Humain-Machine intelligibles permettent la co-évolution des humains et des systèmes
3. Vital d'impliquer les humains (utilisateurs, parties prenantes) dans la conception pour la fiabilité / confiance
4. Généricité et modularité sont importantes pour faire face à la diversité des situations, acteurs, et valeurs

# Objectifs

1. Co-construction de l’éthique entre le système d’IA et les Humains
2. Apprentissage long-terme de comportements éthiques par des systèmes intelligents
3. Borner l’apprentissage des comportements éthiques
4. Évaluation et validation sur des prototypes

# Recherche pluri- et inter-disiciplinaire

- Philosophie des Sciences > Éthique
- Apprentissage Machine > Apprentissage par Renforcement (Multi-Agent / Multi-Objectifs)
- Systèmes Multi-Agent (Normatifs)
- Interaction Humain-Machine

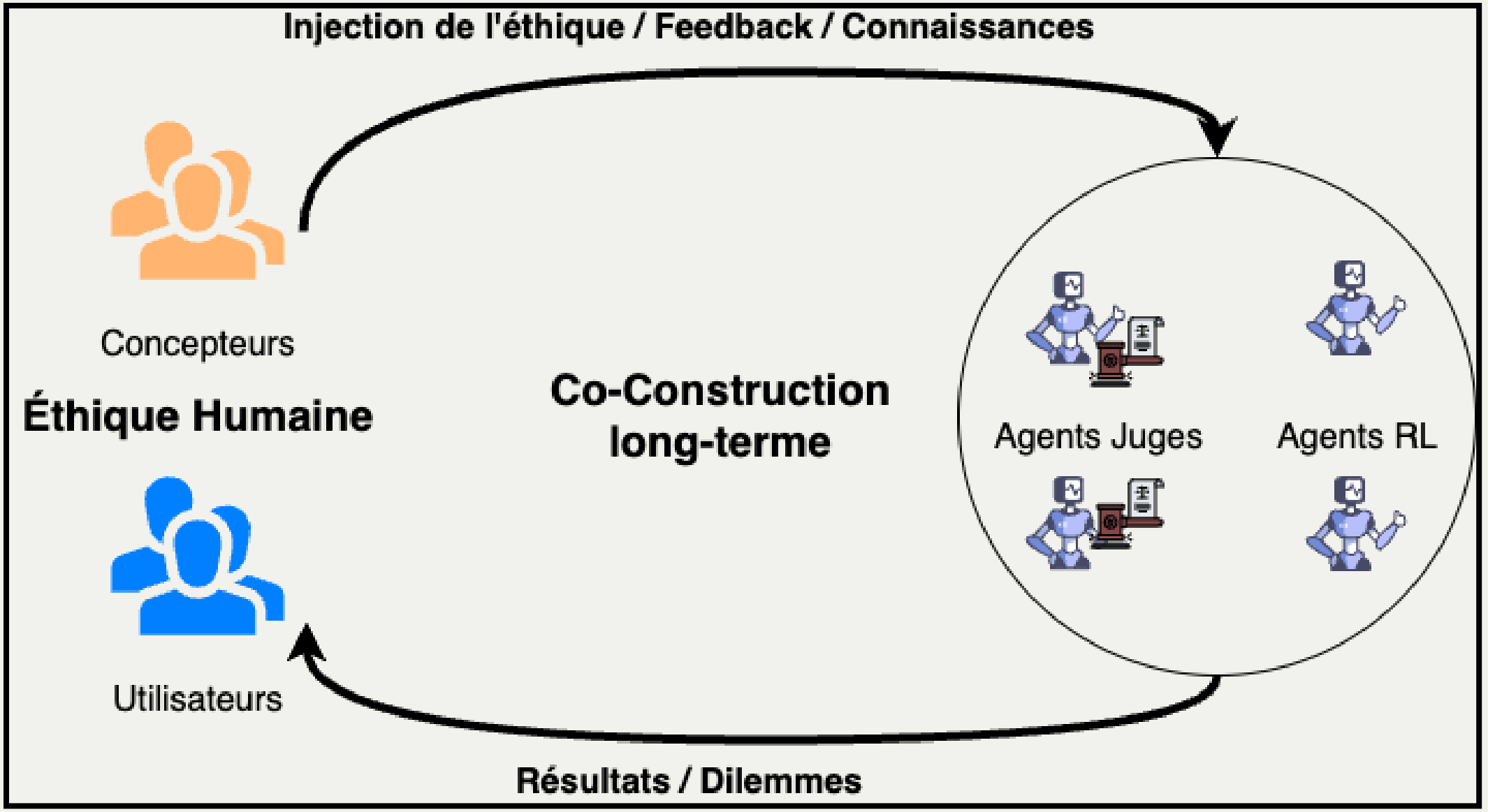
# Co-construction de l’éthique

- Interaction (non-invasive) entre les humains (utilisateurs, concepteurs, ...) et le système
- Système et humains apprennent l’un de l’autre
- 3 processus couplés :
  1. Injection descendante des “considérations éthiques” (valeurs, règles morales, préférences)
  2. Apprentissage à long-terme ascendant
  3. Régulation normative pour encadrer l’apprentissage

# Cas d’application

- Énergie
  - Simulateur d’une *grille intelligente*, répartition de l’énergie parmi multiples agents
  - Conflits entre le confort personnel et l’équité, pas assez d’énergie pour satisfaire tout le monde, écologie
- Transport
  - Choix du mode de transport pour des déplacements dans une métropôle
  - Conflits entre le confort personnel et l’écologie, le confort des habitants (bruit, pollution)

# Résumé


Injection de l'éthique / Feedback / Connaissances
Concepteurs
Éthique Humaine
Co-Construction
long-terme
Agents Juges
Agents RL
Utilisateurs
Résultats / Dilemmes

# Merci de votre attention

Contacts :

- Laetitia Matignon (LIRIS) : Responsable du projet
- Rémy Chaput (LIRIS) : Post-doc